ENHANCING HEVC COMPRESSED VIDEOS WITH A PARTITION-MASKED CONVOLUTIONAL NEURAL NETWORK

Xiaoyi He[1] *Qiang Hu*[1] *Xintong Han*[2] *Xiaoyun Zhang*[1] *Chongyang Zhang*[1] *Weiyao Lin*[1⋆]

[1]Department of Electronic Engineering, Shanghai Jiao Tong University, China
[2]Department of Electrical and Computer Engineering, University of Maryland
(⋆Corresponding Author: wylin@sjtu.edu.cn)

ABSTRACT

In this paper, we propose a partition-masked Convolution Neural Network (CNN) to achieve compressed-video enhancement for the state-of-the-art coding standard, High Efficiency Video Coding (HECV). More precisely, our method utilizes the partition information produced by the encoder to guide the quality enhancement process. In contrast to existing CNN-based approaches, which only take the decoded frame as the input to the CNN, the proposed approach considers the coding unit (CU) size information and combines it with the distorted decoded frame such that the degradation introduced by HEVC is reduced more efficiently. Experimental results show that our approach leads to over 9.76% BD-rate saving on benchmark sequences, which achieves the state-of-the-art performance.

Index Terms— High Efficiency Video Coding, Convolutional neural network, Quality enhancement

1. INTRODUCTION AND RELATED WORK

Recently, the fast development of video capture and display devices has brought a dramatic demand for high definition (HD) contents. High Efficiency Video Coding (HEVC) [1] provides higher compression performance compared to the previous standard H.264/AVC by 50% of bitrate saving on average at a similar perceptual image quality [2]. However, HEVC videos still contain compression artifacts, such as blocking artifacts, ringing effects, blurring, etc.. Therefore, it is desired to study on improving the visual quality of the decoded videos.

Recently, many deep learning based approaches [3, 4, 5, 6] have been proposed to enhance the visual quality of compressed images and videos. [6] designed a CNN to replace the loop filter [7, 8] in HEVC. [3] developed an Artifacts Reduction CNN (ARCNN) built upon [9], which reduces the JPEG compression artifacts. Following [3], [4] and [5] proposed a Variable-filter-size Residual-learning CNN (VR-CNN) and a Quality Enhancement CNN (QECNN) respectively as post-processing techniques to further improve the quality of the compressed videos in HEVC. However, existing works only consider the appearance of input coding units (CUs) or frames, while the partition variations in different CUs and frames are neglected. In practice, since the partition information (e.g., 16×16, 8×8) is introduced by the blockwise processing and quantization of HEVC, this indicates the source of visual compression artifacts. Thus, we use the partition information to effectively guide the quality enhancement process performed by CNN.

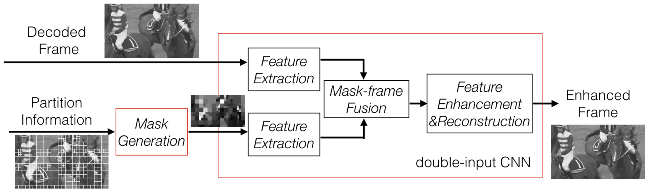
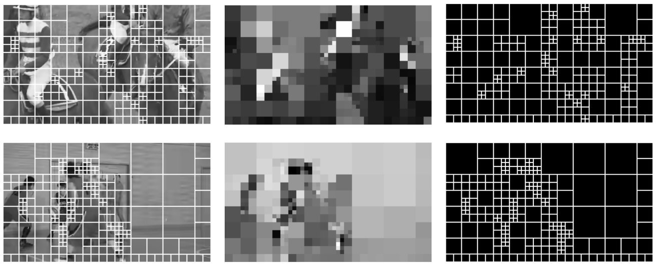


Fig. 1. Overview of the proposed framework.

To this end, we propose a novel approach which first derives a carefully designed mask from a frame's partition information, and then uses it to guide the quality enhancement process of the decoded frame through a double-input CNN. As a result, the visual quality of HEVC-compressed videos can be more properly improved under the same bit rate. The diagram of the proposed approach is shown in Fig. 1. In summary, our contributions are 3 folds:

1. We develop a novel framework that utilizes the partition information to guide the CNN-based quality enhancement process in HEVC, where a mask derived from a decoded frame's partition information is fused with this decoded frame through a double-input CNN to accomplish quality enhancement.
2. Under this framework, we systematically investigate different mask generation and mask-frame fusion methods and find the best strategies. We also demonstrate that our approach is general and can be integrated into the existing HEVC compressed-video enhancement methods to

This work is supported in part by NSFC (61471235), Shanghai 'The Belt and Road' Young Scholar Exchange Grant (17510740100).



(a) original image with partition information (b) Mean-based mask (c) Boundary-based mask

Fig. 2. Two Examples of Boundary-based mask and Mean-based mask.

further improve their performances.

3. We establish a large-scale dataset which contains 202,251 training samples for training reliable compressed-video enhancement models. This dataset will be made publicly available to facilitate further research.

2. OVERVIEW OF OUR APPROACH

The framework of our approach is shown in Fig. 1. Each decoded frame and its corresponding mask, which is generated using the frame's partition information (cf. mask generation in Fig. 1), are fed to a double-input CNN. Inside this CNN, the features of the mask and decoded frame are first extracted through two individual streams and then fused into one (cf. mask-frame fusion). The rest layers of the double-input CNN perform the feature enhancement, mapping, reconstruction, and output the quality-enhanced decoded frame.

3. THE PROPOSED APPROACH

In this section, we first discuss the key components of our approach – mask generation and mask-frame fusion strategies. Then, we describe the proposed double-input CNN.

3.1. Mask generation and mask-frame fusion strategies

Since the block-wise transform and quantization are performed in HEVC during encoding, the quality degradation of compressed frames is highly related to the coding unit splitting. Thus, the partition information contains useful clues for eliminating the artifacts present during the encoding. Considering this, we design a mask based on the partition information of CUs to guide the quality enhancement process.

Generation of the mask. We introduce two strategies to generate masks from an HEVC-encoded frame's partition information:

- Mean-based mask (MM). We fill each partition block in a frame with the mean value of all decoded pixels inside this partition. An example of a generated mean-based mask is shown in Fig. 2b. As we can see that the different partition blocks are properly displayed in the mask. In this way, when we fuse it with the decoded frame during the enhancement process, it can effectively distinguish different partition modes and reduce the compression artifacts more effectively.
- Boundary-based mask (BM). We also introduce a boundary-based mask generation strategy. In this boundary-based mask, the boundary pixels between partitions are filled with value 1 and the rest non-boundary pixels are filled with value 0, as shown in Fig. 2c. The width of the boundary is set to 2.

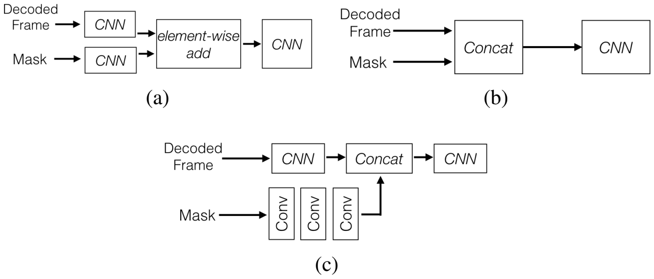


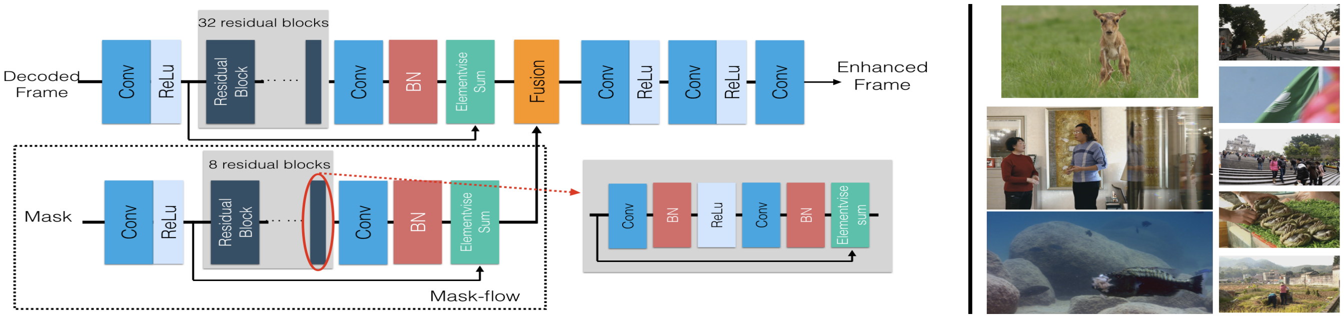
Fig. 3. (a) Add-based fusion. (b) Concatenate-based fusion. (c) Early fusion.

Mask-frame fusion strategies. As we mentioned in Section 2, the mask is fed to CNN and integrated with its corresponding decoded image to get the fused feature maps. We also introduce three strategies to fuse the information of a decoded frame and its corresponding mask:

- Add-based fusion (AF). As shown in Fig. 3a, we first extract the feature maps of the mask using CNN and then combine it with the feature maps of the input frame using element-wise add layer.
- Concatenate-based fusion (CF). We concatenate the mask and frame as the input to the CNN. Then the two-channel image is fed to CNN directly as shown in Fig. 3b.
- Early fusion (EF). We extract the features of mask only using three convolutional layers and integrate it into the network as shown in Fig. 3c.

3.2. Double-input convolutional neural network

The proposed double-input convolutional neural network integrates partition information with add-based fusion strategy and enhances the quality of compressed frames. Its architecture is shown in Fig. 4a. This CNN contains two streams in the feature extracting stage so as to extract features for the decoded frame and its corresponding mask, respectively.



(a)

(b)

Fig. 4. (a) Double-input convolutional neural network with add-based fusion strategy. (b) Example snapshots of our dataset.

Each residual block [10, 11, 12] in the feature extracting stage has two convolutional layers with 3×3 kernels and 64 feature maps, followed by batch-normalization [13] layers and ReLu activation functions. Then, the feature maps of the mask and decoded frame are fused by the add-based fusion strategy and are fed to the rest three convolutional layers. These three layers with 3×3 kernels and 64 feature maps are utilized for feature enhancement, mapping, and reconstruction as described by [3]. When training the network, the Mean Squared Error between the original raw frame and the CNN output is used as the loss function.

Compared with the existing compressed video enhancement methods [4, 5], our network has two differences: (1) We introduce two stream inputs to include both the decoded frame and the partition information. (2) We use a residual architecture to perform the feature extraction. The deep residual stream can capture the feature of input in a more distinctive and stable way.

4. EXPERIMENTAL RESULTS

4.1. Dataset & experimental settings

Dataset. In order to construct a reliable double-input CNN, we establish a large-scale dataset. The dataset is derived from 600 video clips with various resolutions. Fig. 4b shows some snapshots of the video clips. All raw video clips are encoded by HM-16.0 at Low-delay P [14] at QP=22, 27, 32, and 37. In each raw clip and its compressed clip, we randomly select 3 raw frames and the corresponding decoded frames to form 3 training frame pairs. For each frame pair, we divide them into 64×64 sub-images without overlap resulting 202,251 sub-image pairs.

Experimental settings. We implement the proposed model using TensorFlow [15]. During training, we use a mini-batch size of 32. We start with a learning rate of 1e-04, decay the learning rate with a power of 10 at the 20th epochs, and terminate training at 40 epochs. An individual CNN is trained for each QP. In order to save training time, we first train the double-input CNN at QP=37 from scratch and the other networks at QP=32, 27, 22 are fine-tuned from it.

Table 1. Comparison of different mask and fusion methods on ΔPSNR (dB) over HM-16.0 baseline at QP=37

Class	Sequence	1-in	2-in +BM +AF	2-in +MM +EF	2-in +MM +CF	2-in +MM +AF
A	Traffic	0.31	0.37	0.36	0.33	**0.39**
	PeopleOnStreet	0.56	**0.64**	**0.64**	0.61	**0.64**
	Nebuta	0.27	0.20	0.25	0.30	**0.32**
	SteamLocomotive	0.19	0.12	0.18	0.19	**0.22**
B	Kimono	0.36	0.39	0.38	0.39	**0.41**
	ParkScene	0.17	0.19	0.19	0.19	**0.20**
	Cactus	0.23	0.31	0.31	0.27	**0.34**
	BQTerrace	0.18	0.29	0.28	0.29	**0.38**
	BasketballDrive	0.19	0.32	0.31	0.30	**0.35**
C	RaceHorses	0.26	**0.30**	0.29	0.29	0.29
	BQMall	0.10	0.25	0.23	0.27	**0.36**
	PartyScene	0.11	0.17	0.15	0.19	**0.27**
	BasketballDrill	0.22	0.34	0.32	0.32	**0.47**
D	RaceHorses	0.31	**0.42**	0.41	0.41	0.41
	BQSquare	-0.04	0.22	0.16	0.24	**0.50**
	BlowingBubbles	0.13	0.22	0.20	0.22	**0.26**
	BasketballPass	0.19	0.35	0.32	0.36	**0.40**
E	FourPeople	0.44	0.55	0.54	0.53	**0.62**
	Johnny	0.35	0.48	0.47	0.45	**0.54**
	KristenAndSara	0.39	0.56	0.52	0.52	**0.59**
Average		0.25	0.33	0.32	0.33	**0.40**

During the evaluation, we test our trained model on 20 benchmark sequences from the common test conditions of HEVC [16]. The performance of quality enhancement is measured by PSNR improvement (ΔPSNR) and the Rate-distortion performance is measured by the Bjontegaard Distortion-rates (BD-rate) [17] savings over HM-16.0 baseline. Similar to existing works, the performances on Y-channel are evaluated in our experiments.

4.2. Results on different mask generation & mask-frame fusion strategies.

Table 1 compares the performance of different mask generation and mask-frame fusion strategies described in Section

3.1. In Table 1, *1-in* represents a single-input baseline of our approach where the mask-flow input is deleted from the framework of Fig. 4a; *2-in+MM+AF* represents our double-input CNN using the mean-based mask and add-based fusion strategy. Note that the performances of all methods are evaluated by the PSNR gain over HM-16.0 baseline at QP=37. From Table 1, we can have the following observations:

1. When looking at different mask generation strategies, the boundary-based mask strategy (2-in+BM+AF) cannot provide noticeable improvement (0.08 dB over *1-in*). This is because only marking boundary pixels in a mask is less effective in highlighting the partition modes in a frame. Comparatively, the mean-based mask (2-in+MM+AF) can obtain more obvious PSNR improvement (0.15 dB over *1-in*). This indicates its effectiveness in capturing the partition modes in a frame.
2. As for mask-frame fusion strategies, the add-fusion strategy (2-in+MM+AF) can obtain a large PSNR gain of 0.4 dB. This shows the effectiveness of the proposed fusion strategy. Comparatively, the concatenate-fusion (2-in+MM+CF) and early-fusion (2-in+MM+EF) strategies obtains fewer gains. This is probably because these fusion strategies are less compatible with the CNN model used in this paper. Their performances may be more obvious when combined with other CNN models.
3. The best performance is obtained when using mean-based mask and add-fusion (2-in+MM+AF), which can obtain over 0.15 dB improvement over single-input method. This indicates that when strategies are properly selected, introducing partition information is indeed useful to improve the quality of compressed videos.

4.3. Comparison with the existing methods

Table 2 further compares the overall BD-rate saving [17] of different methods over the standard HEVC test model (HM-16.0). Five methods are compared in Table 2: (1) VRCNN [4] which is a benchmark CNN-based compressed-video enhancement method; (2) QECNN-P [5] which is a state-of-the-art compressed-video enhancement method for P frames in HEVC; (3) Our (1-in), which is the single input baseline of our approach; (4) VRCNN+MM+AF, which integrates our partition-mask-based approach into the existing VRCNN method; (5) Our (2-in+MM+AF), which is the full version of our approach with mean-based mask and add-based fusion. Note that in order to have a fair comparison, all methods are trained using the same dataset (i.e., our dataset) and evaluated under the same setting. From Table 2, we can observe that:

1. The full version of our approach (*our+2-in+MM+AF*) achieves the best performance overall the compared methods. Specifically, it can obtain over 9.76% BD-rate reduction from standard HEVC and 4% BD-rate reduction when compared with the state-of-the-art QECNN method. This clearly indicates the effectiveness of our partition-mask-based approach.

Table 2. Comparison of different methods on BD-rate (Y,%) saving over HM-16.0 baseline

Class	Sequence	VRCNN	QECNN-P	Our (1-in)	VRCNN +MM +AF	Our (2-in +MM +AF)
A	Traffic	-6.84	-8.28	-9.271	-9.09	**-11.35**
	PeopleOnStreet	-7.41	-8.66	-9.84	-9.43	**-10.36**
	Nebuta	-5.65	-7.56	-6.23	-6.55	**-7.85**
	SteamLocomotive	-7.71	-9.18	-10.22	-9.89	**-10.6**
B	Kimono	-7.39	-8.70	-9.49	-9.07	**-10.91**
	ParkScene	-3.97	-4.73	-5.4	-5.32	**-6.92**
	Cactus	-5.86	-7.39	-8.13	-8.16	**-10.53**
	BQTerrace	-1.73	-4.87	-7.25	-6.99	**-11.07**
	BasketballDrive	-3.75	-5.91	-6.42	-6.74	**-11.10**
C	RaceHorses	-3.6	-4.78	-5.57	-5.44	**-6.45**
	BQMall	0.11	-2.91	-4.01	-3.97	**-7.62**
	PartyScene	2.72	-1.03	-2.48	-2.08	**-4.84**
	BasketballDrill	-0.08	-2.36	-5.71	-4.64	**-10.65**
D	RaceHorses	-4.05	-5.03	-6.66	-6.41	**-7.58**
	BQSquare	-0.57	-0.11	-2.48	-2.72	**-8.48**
	BlowingBubbles	-0.15	-2.07	-4.12	-3.43	**-6.33**
	BasketballPass	-0.15	-2.37	-4.49	-4.02	**-7.73**
E	FourPeople	-7.12	-9.27	-10.69	-10.33	**-13.91**
	Johnny	-7.00	-9.78	-10.40	-11.41	**-17.22**
	KristenAndSara	-7.13	-9.21	-9.5	-10.56	**-13.78**
Average		-3.81	-5.71	-6.92	-6.81	**-9.76**

2. When integrating our partition-mask strategy, the *VRCNN+MM+AF* can also obtain 3% BD-rate improvement over the original VRCNN method. This demonstrates that our partition-mask-based approach can be easily combined with the existing methods to provide further improved methods.
3. Our baseline single-input method (*our+1-in*) can also obtain satisfactory results when compared with the existing methods (VRCNN, QECNN-P). This implies that the baseline CNN model used in our approach is effective in handling the visual information of the input decoded frames.

5. CONCLUSION

This paper presents a novel approach for enhancing compressed videos in HEVC. Our approach utilizes the partition information already existing in the bitstreams to design a mask and integrate it with the decoded image in CNN to guide the frame quality enhancement process. Experimental results show that our approach is more effective in handling the visual quality degradation introduced by HEVC encoder, and thus obtaining the best post-processing performance. Furthermore, it can also be applied to the existing compressed-video enhancement methods and bring further improvement.

6. REFERENCES

[1] Gary J Sullivan, Jens Ohm, Woo-Jin Han, and Thomas Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on circuits and systems for video technology*, vol. 22, no. 12, pp. 1649–1668, 2012. 1

[2] J. R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand, "Comparison of the coding efficiency of video coding standards including high efficiency video coding (HEVC)," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1669–1684, Dec. 2012. 1

[3] Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang, "Compression artifacts reduction by a deep convolutional network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 576–584. 1, 3

[4] Yuanying Dai, Dong Liu, and Feng Wu, "A convolutional neural network approach for post-processing in HEVC intra coding," in *International Conference on Multimedia Modeling*. Springer, 2017, pp. 28–39. 1, 3, 4

[5] Ren Yang, Mai Xu, Zulin Wang, and Zhenyu Guan, "Enhancing quality for HEVC compressed videos," *arXiv preprint arXiv:1709.06734*, 2017. 1, 3, 4

[6] Woon-Sung Park and Munchurl Kim, "CNN-based in-loop filtering for coding efficiency improvement," in *Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), 2016 IEEE 12th*. IEEE, 2016, pp. 1–5. 1

[7] Andrey Norkin, Gisle Bjontegaard, Arild Fuldseth, Matthias Narroschke, Masaru Ikeda, Kenneth Andersson, Minhua Zhou, and Geert Van der Auwera, "HEVC deblocking filter," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1746–1754, 2012. 1

[8] Chih-Ming Fu, Elena Alshina, Alexander Alshin, Yu-Wen Huang, Ching-Yeh Chen, Chia-Yang Tsai, Chih-Wei Hsu, Shaw-Min Lei, Jeong-Hoon Park, and Woo-Jin Han, "Sample adaptive offset in the HEVC standard," *IEEE Transactions on Circuits and Systems for Video technology*, vol. 22, no. 12, pp. 1755–1764, 2012. 1

[9] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision*. Springer, 2014, pp. 184–199. 1

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 3

[11] Sam Gross and Michael Wilber, "Training and investigating residual nets," *Facebook AI Research, CA.[Online]. Avilable: http://torch.ch/blog/2016/02/04/resnets. html*, 2016. 3

[12] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al., "Photo-realistic single image super-resolution using a generative adversarial network," *arXiv preprint*, 2016. 3

[13] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, 2015, pp. 448–456. 3

[14] Heiko Schwarz, "Hierarchical b pictures," *Joint Video Team (JVT) of ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q. 6, JVT-P014*, 2005. 3

[15] Martín Abadi, Ashish Agarwal, and Paul Barham et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, Software available from tensorflow.org. 3

[16] F Bossen and HM Common, "test conditions and software reference configurations, jct-vc doc," *L1100, Jan*, 2013. 3

[17] Gisle Bjontegaard, "Calculation of average psnr differences between rd-curves," in *ITU-T Q. 6/SG16 VCEG, 15th Meeting, Austin, Texas, USA, April, 2001*, 2001. 3, 4